Systematic bias in transport model calibration arising from the variability of linear data projection

Wai Wong and S.C. Wong

Department of Civil Engineering, The University of Hong Kong, Pokfulam, Hong Kong

Abstract

In transportation and traffic planning studies, accurate traffic data are required for reliable model calibration to accurately predict transportation system performance and ensure better traffic planning. However, it is impractical to gather data from an entire population for such estimations because the widely used loop detectors and other more advanced wireless sensors may be limited by various factors. Thus, making data inferences based on smaller populations is generally inevitable. Linear data projection is a commonly and intuitively adopted method for inferring population traffic characteristics. It projects a sample of observable traffic quantities such as traffic count based on a set of scaling factors. However, scaling factors are subject to different types of variability such as spatial variability. Models calibrated based on linearly projected data that do not account for variability may introduce a systematic bias into their parameters. Such a bias is surprisingly often ignored. This paper reveals the existence of a systematic bias in model calibration caused by variability in the linear data projection. A generalized multivariate polynomial model is applied to examine the effect of this variability on model parameters. Adjustment factors are derived and methods are proposed for detecting and removing the embedded systematic bias. A simulation is used to demonstrate the effectiveness of the proposed method. To illustrate the applicability of the method, case studies are conducted using real-world global positioning system data obtained from taxis. These data calibrate the Macroscopic Bureau of Public Road function for six 1x1 km regions in Hong Kong.

Keywords: Systematic Bias; Model Calibration; Linear Data Projection; Macroscopic Bureau of Public Road; GPS

1. Introduction

Reliable model calibration is crucial in transportation studies as it helps to establish a better understanding of the interactions between transportation infrastructure, vehicles and road users. Accurate model calibration leads to better urban and traffic planning and the implementation of traffic management and control measures. Consequently, it helps to develop a less congested and more efficient network, keeps a city more economically competitive and decreases traffic emissions. In addition, due to the irreversible patterns of development restricted by infrastructures and the critical role of infrastructure in promoting economic growth (Carlsson et al., 2013), careful planning with the support of reliable model calibration is essential for preventing the misuse of the public budget and resources.

The accurate measurement and estimation of traffic quantities result in reliable model calibration. Technological advancements have improved the accuracy and efficiency of traffic data collection methods over the past decades. Hand tally measurement has gradually been replaced by automatic systems such as inductive loop sensors, radar and television cameras. In addition to point measurement, methods for measuring along a length of road and the collection of data by a moving observer have also been developed. The rapid development of intelligent transportation systems has made it possible to conduct measurements over a wide area at a relatively low cost.

On-road fixed detectors such as inductive loop sensors are still the most commonly adopted means of collecting traffic data for important roadways, as such methods provide an acceptable level of accuracy with minimal effort. However, high installation and maintenance costs sometimes make it impractical or economically unviable to ubiquitously deploy these sensors on all highways and the entire arterial network (Herrera and Bayen, 2010, Herrera et al., 2010). Hence, the coverage is normally limited to a subset of links (Caceres et al., 2012).

Given that vehicle movement can be interrupted by signals, the travel time estimates of loop detectors could be inaccurate. In principle, a vehicle re-identification system can improve the accuracy as follows. Sensors installed at the two ends of a selected arterial link record the times when a vehicle passes by and measure its signature. The travel time of the vehicle is calculated when the signature is matched at the two consecutive locations of the link (Kwong et al., 2009). The radio frequency identification (RFID) transponders (Wright and Dahlgren, 2001, Ban et al., 2010), license plate recognition (LPR) systems (Herrera et al., 2010) and other unique tags are readily available utilities for this scheme. However, in addition to raising privacy concerns, these systems are similarly limited by the cost of sensor deployment over the entire arterial network, thus restricting coverage. Kwong et al. (2009) presented a scheme based on matching signatures measured by wireless magnetic sensors installed at the two ends of the arterial link. Although this scheme is able to avoid the risk of privacy issues, it fails to resolve cost and coverage problems. More recently, the Bluetooth Media Access Control Scanner (BMS) was proposed as a complementary traffic data source (Bhaskar and Chung, 2013). However, Jie et al. (2011) identified the poor quality of its data and the uncertainty surrounding its identification of Bluetooth device carriers (i.e., whether a carrier belongs to a vehicle, a cyclist or a pedestrian).

Cellular systems were introduced a decade ago (Bolla and Davoli, 2000, Ygnace and Drane, 2001, Zhao, 2000) to overcome the limitations imposed by expensive implementation costs and the limited coverage of stationary roadside equipment (Herrera et al., 2010) in systems such as loop detectors and vehicle re-identification systems. However, because the use of cell phones while driving disrupts drivers' attention (Liang et al., 2007), it is prohibited or discouraged in many countries, thus limiting the application of the proposed models. Moreover, flow measurements from cellular systems follow an aggregate format for each group of links intercepting the corresponding inter-cell boundary (Caceres et al., 2012), making it impossible to estimate traffic flow for any individual link.

Advancements in global positioning systems (GPSs) have made it possible to collect data from GPS-equipped vehicles. These systems have been widely adopted to extend the coverage of data collected from stationary roadside equipment to almost the entire network at a relatively low cost (Miwa et al., 2013). Many recent travel time estimation studies have been based on GPS probe vehicle data (Nanthawichit et al., 2003, Hofleitner et al., 2012, Peer et al., 2013, Herring et al., 2010, Jenelius and Koutsopoulos, 2013, Zheng and Van Zuylen, 2013, Zhan et al., 2013). Although they lend potential to future global coverage, these probe vehicle data come from various sources that present specific challenges. First, fleet data (FedEx, UPS, taxis, etc.) (Moore et al., 2001, Schwarzenegger et al., 2008, Bertini and Tantiyanugulchai, 2004; Wong et al., 2014) pose bias problems due to the operational constraints and specific travel patterns involved. Second, participatory sensing data taken from industry models (INRIX, Waze, etc.) are unpredictable, and no single company has ubiquitous coverage (Hofleitner et al., 2012). Moreover, the added cost of equipping every vehicle with GPS trackers coupled with potential privacy issues prevent this system from being applied on a global scale, making direct measurement of total traffic flows implausible.

Despite the advancement of technologies, the collection of traffic data via different devices remains limited by various factors. Mathematical techniques used for traffic data estimations, such as sampling methods, filtering algorithms and data scaling, offer possible solutions to the problems presented by data acquisition. Linear data projection is a prevalent data scaling method that infers population traffic characteristics by projecting the observable traffic characteristics of a smaller population via the mean of a set of scaling factors.

The scaling factors used in linear data projections vary by situation. Example scaling factors include traffic composition ratios and passenger car units (PCUs). The factor is usually a random variable that is subject to variability and assumed to follow a distribution, rather than a constant. Depending on the sampling method used, the variance of the sampled scaling factor measures different types of variability, such as spatial and temporal variability. If traffic composition ratios are sampled across a network, then the variance measures spatial variability. Contrary to the usual assumption, a PCU is not essentially static (Chandra et al., 1995). Thus, if it is selected as the scaling factor, its variance during different time points at the same site measures temporal variability. Because the mean of the distribution is the most probable observed scaling factor, it is usually adopted in linear data projections.

Linear data projections are especially useful for traffic data estimations in situations where direct measurement is not possible such as the lack of spatial coverage of sensors. For instance, a linear data projection can be adopted to estimate an hourly total traffic flow on a link where on-road fixed detectors are not installed. Assuming that occupied taxi flow is observable on every roadway in a network and that total traffic flow is only observable on a subset of links outfitted with detectors in the network, the total traffic-to-occupied-taxi ratio can be the chosen scaling factor, and is assumed to follow a distribution over a region due to geographical proximity. Scaling factors can be sampled at sites outfitted with detectors. The mean of the sampled scaling factors is the expected total traffic-to-occupied-taxi ratio across that region in the long run. The variance of the sampled scaling factors measures the spatial variability of the total traffic-to-occupied-taxi ratio within this network. If the hourly

occupied taxi flow on the link of interest is 10 veh/h and the mean of scaling factors sampled at the nearby sites is 100, the hourly total traffic on this link can be estimated by the product of the mean of the scaling factors and the occupied taxi flow, which is 1000 veh/h in this case. In their study of urban-scale macroscopic fundamental diagrams, Geroliminis and Daganzo (2008) leveraged the notion of linear data projection to infer the total traffic flow of sites without loop detector installations from the flow of a small group of GPS-equipped taxis, using the traffic composition ratio as the scaling factor. This scaling method is not limited to projecting traffic flow. It can also be used to infer other quantities such as trip completion rates, vehicular accumulations and space-mean speeds (Geroliminis and Daganzo, 2008).

Due to its simple concept, linear data projection has been widely adopted in many real-world situations that necessitate data scaling via scaling factor. However, scaling factors such as traffic composition ratios and PCUs are random variables with variations rather than absolute constants. Systematic bias may be embedded in the parameters of a model calibrated based on linearly projected data because the variance, skewness, kurtosis and even higher-ordered moment of the distribution of the scaling factor are not captured in the linear data projection.

This embedded systematic bias remains unexplored in the field, as it is not easily evident. To reveal and demonstrate the existence of the bias, a numerical example of the calibration of a simple polynomial model simulating a linear data projection is presented as follows:

$$y = a + b\,X^n = a + b\,(fx)^n$$

where x is the observable independent variable; f is the scaling factor of x; $X = fx$ is the projected value; y is the observable dependent variable; and a, b and n are the model parameters.

Ten thousand data points of x, which serve as the observed data for the independent variable, are sampled from a uniform distribution with a domain from 0 to 1. Because scaling factors are generally positive, a lognormal distribution with $\bar{f} = 1$ and $\sigma_f = 0.2$ is chosen to sample the corresponding scaling factors for the 10,000 samples of x. $\bar{f}$ and σ_f are respectively the mean and standard deviation of the scaling factor f. Depending on the sampling method used, both the standard deviation σ_f and variance ${\sigma_f}^2$ can measure variability such as the spatial variation or temporal variation of the scaling factors across the dimension under consideration. Assuming that $a = 1$, $b = 1$ and $n = 3$, the corresponding 10,000 points of y and $X = fx$, which serve as the observed data for the dependent and projected independent variable, can be calculated based on the assumed values of the parameters and sampled x and f.

Suppose that the values of all individual X are no longer available and can only be estimated via a linear projection function based on the mean value of f, $\bar{f}$, a common real-world occurrence. Regression analysis is then conducted between y and the linearly projected X. The calibrated values of the parameters are $\hat{a} = 0.999$ and $\hat{b} = 1.130$. It is obvious that $\hat{a}$

is close to the assumed true value. However, the calibrated value of b is apparently overestimated. The overestimation of $\hat{b}$ (+13.0%) reveals the existence of systematic bias due to the ignorance of scaling factor variability in the linear data projection. A linear data projection provides good estimates of unobservable independent variables because it captures the first moment of the scaling factor that carries most of the information. However, such point estimates are not sufficient for reliable model calibration.

Models depicting the characteristics and performance of a network use fundamental diagrams and both link- and area-based cost-flow functions. These models such as volume delay functions (e.g., Spiess, 1990, Akcelik, 1978, Tisato, 1991, Davidson, 1966, Akçelik, 1980) and speed-density relationships (e.g., Jayakrishnan et al., 1995, Kerner and Konhäuser, 1994, Drake et al., 1967, Drew, 1965, Munjal and Pipes, 1971, Pipes, 1967, MacNicholas and Board, 2008, Del Castillo and Benitez, 1995a, Del Castillo and Benitez, 1995b, Van Aerde, 1995) require traffic speed, flow and density data, the three most important quantities in transportation. However, if a non-negligible subset of links within a network is not equipped with adequate instruments for direct traffic data measurement, which is usually the case in urban transportation network (Lederman and Wynter, 2011), a linear data projection may be leveraged using the observable traffic data of a smaller population to estimate traffic data. Models calibrated based on these linearly projected data may be systematically biased. To remove this bias, information provided by the scaling factor variability should be incorporated into the calibration of the model.

This paper fills the aforementioned knowledge gap by proposing the incorporation of adjustment factors that capture scaling factor variability into the model calibration process. We derive global adjustment factors that correct the calibrated sensitivity parameters of chosen generalized multivariate models in polynomial form. The Bureau of Public Roads (BPR) function adopted in the Highway Capacity Manual (Transportation Research Board, 2000) is a polynomial function that can model the relationship between travel time and the traffic volume in a link. It is commonly used in many European countries and the United States (Dowling et al., 1998, Lum et al., 1998) and plays an important role in static user equilibrium analysis (García-Ródenas and Verastegui-Rayo, 2013). The case studies section presents calibrations of Macroscopic Bureau of Public Roads (MBPR) functions using real-life GPS data and demonstrates the application of the derived global adjustment factor. The main contribution of the proposed global adjustment factor is that it can remove the systematic bias introduced in the calibrated parameters and hence ensure more accurate model calibration.

The remainder of this paper is structured as follows. In Section 2, the existence of the systematic bias embedded in parameters calibrated from linear projected data is proven based on a Taylor series expansion. In Section 3, the adjustment factor for models in generalized multivariate polynomial form is derived. The metric measuring the extent of the systematic bias, factors affecting the extent of the embedment of the systematic bias and the method for removing the bias embedded in the calibrated sensitivity parameters are also presented in Section 3. Section 4 presents a simulation to illustrate the significant correction power of the derived global adjustment factors for multivariate functional models, and demonstrates that

the applicability of the global adjustment factor is not restricted to the magnitudes of the mean and coefficient of variation (CV) of the scaling factor. In Section 5, real-world taxi GPS data are used to calibrate the macroscopic cost-flow function, and the derived global adjustment factor is applied in an illustrative case study. Finally, Section 6 summarizes the findings of the paper and discusses possible future research directions.

2. Existence of systematic bias

This section reveals the necessary and sufficient condition for the introduction of systematic bias into the calibrated model parameters arising from linearly projected data, and thereby proves its existence. The origin of the systematic bias is then discussed.

2.1 Necessary and sufficient condition for the introduction of systematic bias

Consider a function $y = G(z)$ of any form, where z is constituted by the sum of the products of a set of scaling factors and a set of observable independent variables, i.e., $z = \sum_{i=1}^{m} f_i x_i$; x_i is the observable independent variable; f_i is the scaling factor of x_i, which is assumed to follow any distribution with mean $\bar{f}$ and variance ${\sigma_f}^2$; and m is the number of terms used to construct the quantity, z.

In most cases, it is impossible or impractically expensive and labor-intensive to collect data for z compared with x_i. In practice, data for the observable variable, x_i, can be collected in relatively cheaper ways. It is assumed that the scaling factor, f_i, of each individual x_i follows a distribution. In theory, the scaling factor, f_i, can be assumed to follow any distribution. However, the chosen distribution depends on the properties of the scaling factor. For instance, if the scaling factor is a non-negative random variable with a lower relative frequency at high values, then lognormal distribution is an assumed candidate distribution. The first and second moments of the distribution can be estimated from another set of scaling factors collected from another source under similar conditions. Each set of observable variable, x_i, can be scaled by the estimated mean of the scaling factor, $\bar{f}$, i.e., $\sum_{i=1}^{m} \bar{f} x_i$, as an estimate of the target variable, z. The calibrated model based on the linearly projected data is $G(\bar{\boldsymbol{f}})$, which may be a model calibrated with systematically biased parameters. Proposition 1 states the necessary and sufficient condition for the introduction of systematic bias. In other words, if the following conditions (the model to be calibrated is a non-linear function of the scaling factor and the scaling factor is subject to variability) are not satisfied, the calibrated model is unbiased even the linear data projection is employed.

Proposition 1:
Systematic bias is embedded in the calibrated parameters of models calibrated from linearly projected data, regardless of the distribution of the scaling factor and the form of the model $G(z)$, as long as it is a non-linear function of the scaling factor and the scaling factor is subject to variability.

Proof: Approximate y by a Taylor series expansion with the center at $f_i = \bar{f}, \forall i \in \mathbb{N}^+$

$$y = G(\bar{\boldsymbol{f}}) + \sum_{i=1}^{m} \frac{\partial G(\bar{\boldsymbol{f}})}{\partial f_i}(f_i - \bar{f}) + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \frac{\partial^2 G(\bar{\boldsymbol{f}})}{\partial f_i f_j}(f_i - \bar{f})(f_j - \bar{f}) + \cdots \qquad (1)$$

Ignoring higher order terms and taking the expectation on both sides,

$$E(y) = G(\bar{\boldsymbol{f}}) + \sum_{i=1}^{m} \frac{\partial G(\bar{\boldsymbol{f}})}{\partial f_i} E(f_i - \bar{f}) + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \frac{\partial^2 G(\bar{\boldsymbol{f}})}{\partial f_i f_j} E[(f_i - \bar{f})(f_j - \bar{f})] \qquad (2)$$

where $E(f_i) = \bar{f}$, $E(f_i - \bar{f}) = 0$. Assuming that $f_i, \forall i \in \mathbb{N}^+$ are independent of each other, $E[(f_i - \bar{f})(f_j - \bar{f})] = 0,\ \forall i, j \in \mathbb{N}^+ \backslash [i = j]$ and $E\left[(f_i - \bar{f})^2\right] = {\sigma_f}^2$. It follows that

$$E(y) = G(\bar{\boldsymbol{f}}) + \frac{1}{2}{\sigma_f}^2 \sum_{i=1}^{m} \frac{\partial^2 G(\bar{\boldsymbol{f}})}{\partial {f_i}^2} \qquad (3)$$

Thus, both the mean of the scaling factor, $\bar{f}$, and the variance of the scaling factor, ${\sigma_f}^2$, contribute to the expectation of y.

$$E(y) \neq G(\bar{\boldsymbol{f}})$$

$$\Leftrightarrow \qquad \frac{1}{2}{\sigma_f}^2 \sum_{i=1}^{m} \frac{\partial^2 G(\bar{\boldsymbol{f}})}{\partial {f_i}^2} \neq 0$$

$$\Leftrightarrow \qquad {\sigma_f}^2 \neq 0 \qquad and \qquad \sum_{i=1}^{m} \frac{\partial^2 G(\bar{\boldsymbol{f}})}{\partial {f_i}^2} \neq 0$$

■

Systematic bias may be introduced if $G(z)$ is a nonlinear function of the scaling factor. However, $G(z)$ is unrestrictive to any kind of model form. In particular, according to Equation (3), if $G(\bar{\boldsymbol{f}})$ is a factor of $\sum_{i=1}^{m} \frac{\partial^2 G(\bar{\boldsymbol{f}})}{\partial {f_i}^2}$, then the variance of the scaling factor can be easily grouped with the model parameters. The model parameters affected by the variance of the scaling factor and the effect of the variance on those parameters can be easily identified. Examples of such models are generalized multivariate polynomial models, the exponential function and some trigonometric functions such as the sine and cosine functions. This paper examines the generalized multivariate polynomial model. If $G(\bar{\boldsymbol{f}})$ is not a factor of

$\sum_{i=1}^{m} \frac{\partial^2 G(\bar{f})}{\partial {f_i}^2}$, then the effect of the scaling factor variance on the model parameters is fuzzy, as the variance of the scaling factor cannot be easily fused with the model parameter.

2.2 Origin of systematic bias

Embedded systematic bias stems from uncaptured scaling factor variability. Equation (3) demonstrates that the second moment of a scaling factor distribution also plays a role in contributing the mean of the response variable. Direct model calibration based on linearly projected data implies an ignorance of the information provided by the second moment. The discrepancy between the true model, $E(y)$ or $G(z)$, and the calibrated model, $G(\bar{\boldsymbol{f}})$, is the second term on the right-hand side of Equation (3), and undesirably remains in the error component. This unexplained and hidden contribution to the mean value of the response variable, which is a function of the independent variable, is then captured by the error distribution mean. The violation of the mean zero-error distribution results in a systematically biased calibrated model.

Adopting linear data projection, i.e., replacing each individual scaling factor, f_i, in $\sum_{i=1}^{m} f_i x_i$ with $\bar{f}$, leads to translational movements and the systematic scattering of the data points of the true model along the independent variable space. This data point distortion is equivalent to shifting the entire distribution of data points at each observation along the dependent variable space, such that the most probable observed value differs from the true value by the exact value of the second term in Equation (3). Linear data projection generates systematically distorted data points. Furthermore, calibrating a model with systematically distorted data points naturally creates a systematically biased model.

3. Global adjustment factors for generalized multivariate polynomial models

The paper uses a generalized multivariate polynomial model to examine the effect of the ignorance of scaling factor variability in model calibration. The goal of this section is to derive the global adjustment factors that capture scaling factor variability. A metric measuring the extent of the embedment of systematic bias is proposed, and the factors affecting the amount of introduced systematic bias are discussed. A method for incorporating the captured variability to correct the calibrated parameters is then introduced.

3.1 Derivation of global adjustment factors

Consider the following model in polynomial form with n + 1 terms:

$$y = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n \tag{4}$$

where $z = \sum_{i=1}^{m} f_i x_i$; a_0, a_1, a_2, …,a_{n-1} and a_n are the parameters to be calibrated.

Thus, the preceding model can be generalized in a multivariate functional form as follows:

$$y = a_0 + a_1\left(\sum_{i=1}^{m} f_i x_i\right) + a_2\left(\sum_{i=1}^{m} f_i x_i\right)^2 + \cdots + a_n\left(\sum_{i=1}^{m} f_i x_i\right)^n \tag{5}$$

Proposition 2 stated below presents a global adjustment factor for the calibrated sensitivity parameter, $\hat{a}_k$, of the k^{th} term with exponent k. Each calibrated sensitivity parameter of the generalized multivariate polynomial function has its own global adjustment factor and can be corrected independently. The global adjustment factor is expressed in terms of the scaling factor variance, and hence is able to capture the lost variability information in linear data projection.

Proposition 2:
The global adjustment factor, $\bar{F}_k$, for the calibrated sensitivity parameter, $\hat{a}_k$, of the k^{th} term with exponent k is given by

$$\bar{F}_k = 1 + \left[\frac{k(k-1)}{2}\right]\left[\left(\frac{\sigma_f}{\bar{f}}\right)^2\right]\left[\frac{\sum_{\iota=1}^{m} x_\iota^{\,2}}{\left(\sum_{\iota=1}^{m} x_\iota\right)^2}\right]$$

Proof: Consider the k^{th} term with exponent k,$T_k = a_k(\sum_{i=1}^{m} f_i x_i)^k$:

$$\frac{\partial T_k}{\partial f_j} = a_k k\left(\sum_{i=1}^{m} f_i x_i\right)^{k-1} x_j$$

$$\frac{\partial^2 T_k}{\partial {f_j}^2} = a_k k(k-1)\left(\sum_{i=1}^{m} f_i x_i\right)^{k-2} {x_j}^2$$

Using Equation (3),

$$E(y) = a_0 + a_1\left(\sum_{i=1}^{m} \bar{f} x_i\right) + a_2\left[1 + \frac{(2\cdot 1){\sigma_f}^2 \sum_{i=1}^{m} {x_i}^2}{2\left(\sum_{i=1}^{m} \bar{f} x_i\right)^2}\right]\left(\sum_{i=1}^{m} \bar{f} x_i\right)^2 + \cdots$$

$$+ a_n\left[1 + \frac{n(n-1){\sigma_f}^2 \sum_{i=1}^{m} {x_i}^2}{2\left(\sum_{i=1}^{m} \bar{f} x_i\right)^2}\right]\left(\sum_{i=1}^{m} \bar{f} x_i\right)^n$$

Hence, $\forall k \in \mathbb{N}^0$, and

$$E(y) = \sum_{k=0}^{n} a_k \left[1 + \frac{k(k-1)\sigma_f{}^2 \sum_{i=1}^{m} x_i{}^2}{2(\sum_{i=1}^{m} \bar{f} x_i)^2}\right] \left(\sum_{i=1}^{m} \bar{f} x_i\right)^k \quad (6)$$

Again, the mean value of y is dependent on both the mean and variance of the scaling factor. However, in practice, the collected data of the observable independent variable are normally linearly projected using the mean of the scaling factor, and the information provided by the scaling factor variance is ignored in the model calibration. Thus, systematic biases are introduced into some of the sensitivity parameters.

In Equation (6), the variance of the scaling factors is taken out from each scaling factor, f_i, and absorbed by the sensitivity parameters of each term. Comparing Equation (6) with the original model in Equation (5), we can define the local adjustment factor, F_k, for the calibrated sensitivity parameter, $\hat{a}_k$, of the k^{th} term with exponent k as follows:

$$F_k = 1 + \frac{k(k-1)\sigma_f{}^2 \sum_{i=1}^{m} x_i{}^2}{2(\sum_{i=1}^{m} \bar{f} x_i)^2} \quad (7)$$

or

$$F_k = 1 + \left[\frac{k(k-1)}{2}\right]\left[(\frac{\sigma_f}{\bar{f}})^2\right]\left[\frac{\sum_{i=1}^{m} x_i{}^2}{(\sum_{i=1}^{m} x_i)^2}\right] \quad (8)$$

The adjustment factor for each calibrated sensitivity parameter is dependent on the exponent, k, the CV of the scaling factor, $\frac{\sigma_f}{\bar{f}}$, and $\frac{\sum_{i=1}^{m} x_i{}^2}{(\sum_{i=1}^{m} x_i)^2}$, where x_i are the collected observable independent variables. The adjustment factor is localized because $\frac{\sum_{i=1}^{m} x_i{}^2}{(\sum_{i=1}^{m} x_i)^2}$ is dependent on each set of collected observable independent variables, $x_1, x_2, \ldots, x_m$.

According to the Cauchy-Schwarz inequality, $\forall x_i, y_i \geq 0$,

$$0 \leq \left(\sum_{i=1}^{m} x_i y_i\right)^2 \leq \left(\sum_{i=1}^{m} x_i{}^2\right)\left(\sum_{i=1}^{m} y_i{}^2\right)$$

Setting all $y_i = 1$,

$$0 \leq \left(\sum_{i=1}^{m} x_i\right)^2 \leq \left(\sum_{i=1}^{m} x_i{}^2\right) m$$

Given that $\sum_i^m x_i \neq 0$,

$$\frac{1}{m} \leq \frac{\sum_{i=1}^{m} {x_i}^2}{(\sum_{i=1}^{m} x_i)^2}$$

The equality of the lower bound holds when $x_1 = x_2 = \cdots = x_m$. As $m \to \infty$, the lower bound $1/m \to 0$. Furthermore,

$$\left(\sum_{i=1}^{m} x_i\right)^2 = \sum_{i=1}^{m} {x_i}^2 + 2\sum_{i<j}^{m} x_i x_j$$

$$\geq \sum_{i=1}^{m} {x_i}^2$$

Thus,

$$\frac{1}{m} \leq \frac{\sum_{i=1}^{m} {x_i}^2}{\left(\sum_{i=1}^{m} x_i\right)^2} \leq 1 \tag{9}$$

The equality of the upper bound holds when all of the observed independent variables but one are equal to zero.

In general, as m increases, $\frac{\sum_{i=1}^{m} {x_i}^2}{\left(\sum_{i=1}^{m} x_i\right)^2}$ decreases, and the adjustment factor moves closer to 1. In other words, the systematic bias decreases because the random effects among different scaling factors cancel one another out as m increases. In the limiting case,

$$\lim_{m\to\infty} \frac{\sum_{i=1}^{m} {x_i}^2}{\left(\sum_{i=1}^{m} x_i\right)^2} = 0 \tag{10}$$

Hence, the adjustment factor, F_k, tends to 1. In such a case, no adjustment is necessary.

A local adjustment factor is determined by each set of collected observable independent variables, $x_1, x_2, \ldots, x_m$. A distribution of $\frac{\sum_{i=1}^{m} {x_i}^2}{\left(\sum_{i=1}^{m} x_i\right)^2}$ with a lower bound = $\frac{1}{m}$ and an upper bound = 1 is formed by all of the sets of collected observable independent variables. In practical terms, $\frac{\sum_{i=1}^{m} {x_i}^2}{\left(\sum_{i=1}^{m} x_i\right)^2}$ varies across different observations. In this paper, this effect is represented using the mean value of all of the individual observations, $\overline{\frac{\sum_{i=1}^{m} {x_i}^2}{\left(\sum_{i=1}^{m} x_i\right)^2}}$, based on which a global adjustment factor, $\bar{F}_k$, is defined for the calibrated sensitivity parameter, $\hat{a}_k$, of the k^{th} term with exponent k:

$$\bar{F}_k = 1 + \left[\frac{k(k-1)}{2}\right]\left[\left(\frac{\sigma_f}{\bar{f}}\right)^2\right]\left[\frac{\sum_{i=1}^{m} x_i^2}{\left(\sum_{i=1}^{m} x_i\right)^2}\right] \quad (11)$$

■

Note that when m decreases to 1, the multivariate polynomial function decreases to a polynomial function of a single variable. The global adjustment factor for each sensitivity parameter is exactly the same as the local adjustment factor, as $\frac{\sum_{i=1}^{m} x_i^2}{\left(\sum_{i=1}^{m} x_i\right)^2} = 1$.

The derived global adjustment factors capture the scaling factor variability information, which is normally ignored in model calibration and can be used to remove the systematic bias as shown in Section 3.3.

3.2 Extent of embedded systematic bias

The global adjustment factor captures the ignored scaling factor variability. Hence, it contains information related to the extent of the embedded systematic bias it causes. The percentage of embedded systematic bias in the calibrated sensitivity parameter, $\hat{a}_k$, is defined as

$$B(\hat{a}_k) = (\bar{F}_k - 1) \times 100\% \quad (12)$$

where $B(\hat{a}_k)$ is the percentage of embedded systematic bias in the calibrated parameter, $\hat{a}_k$.

A zero value of $B(\hat{a}_k)$ indicates an unbiased calibrated parameter. If it is positive, overestimation of the calibrated parameter is anticipated. Underestimation of the calibrated parameter is expected when $B(\hat{a}_k)$ is negative in value.

The extent of the embedded systematic bias is governed by the exponent, k, the CV of the scaling factor, $\frac{\sigma_f}{\bar{f}}$, and $\frac{\sum_{i=1}^{m} x_i^2}{\left(\sum_{i=1}^{m} x_i\right)^2}$. For a given value of $\frac{\sum_{i=1}^{m} x_i^2}{\left(\sum_{i=1}^{m} x_i\right)^2}$, e.g., 0.5, Figure 1 shows the relationship between the extent of the embedded systematic bias and the CV of the scaling factor, ranging from 0 to 1, at different exponents: $k = 0$, $k = 0.5$, $k = 1$, $k = 2$, $k = 3$ and $k = 4$. Figure 2 reveals the relationship between the extent of the embedded systematic bias and the exponent, which ranges from 0 to 2, at different CVs: $(\sigma_f/\bar{f})^2 = 0$, $(\sigma_f/\bar{f})^2 = 0.5$ and $(\sigma_f/\bar{f})^2 = 1.0$.

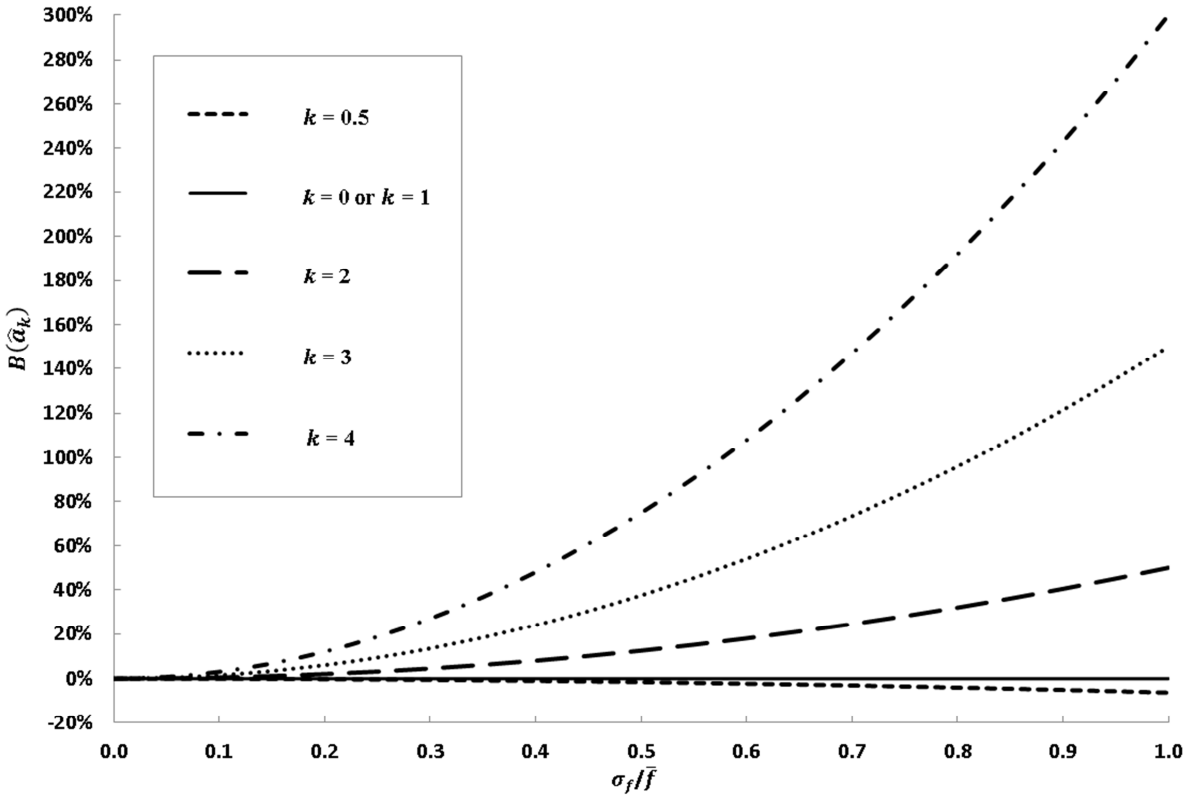

Figure 1. Variation of the global adjustment factor against the CV of the scaling factor.

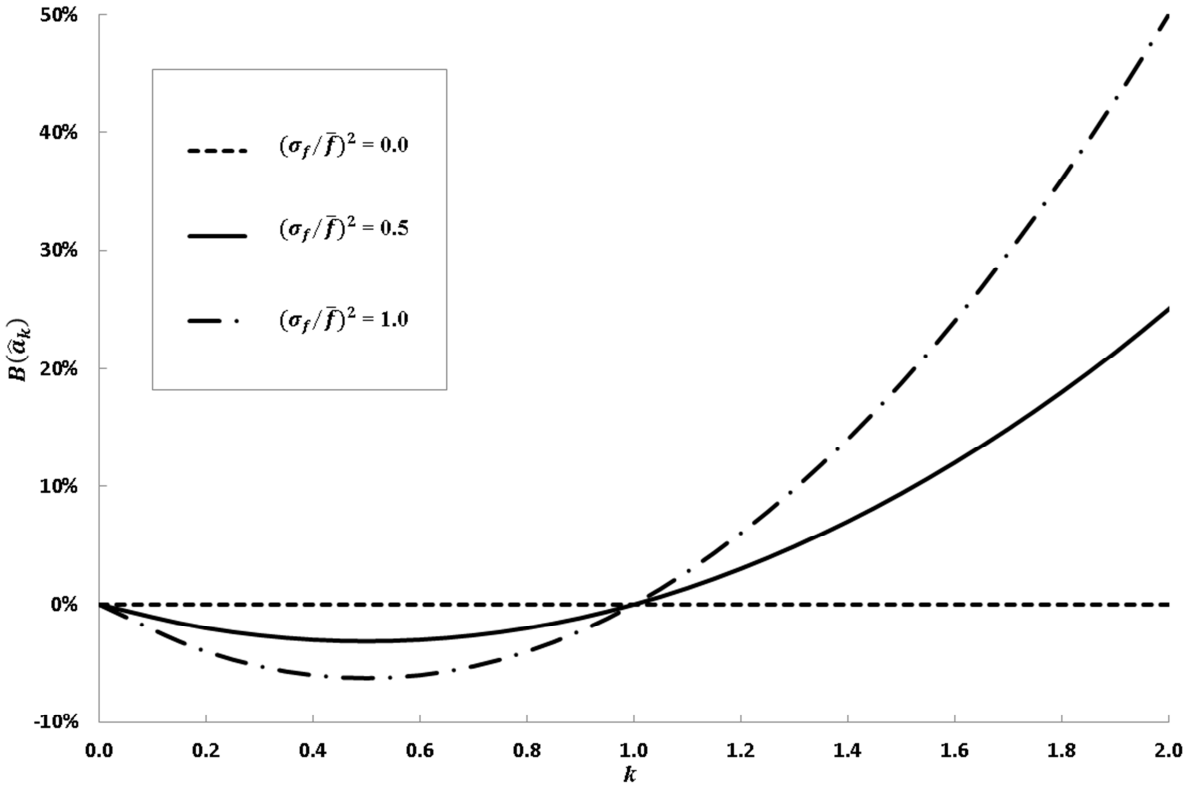

Figure 2. Variation of the global adjustment factor against the exponent.

When the CV of the scaling factor is 0, suggesting zero variability in the scaling factor, no systematic bias is introduced into the calibrated parameter regardless of the value of the exponent, as shown in Figure 1, because the scaling factor for scaling the observable independent variable is exact without any loss of information. Except for $k = 0$ and $k = 1$, the extent of embedded systematic bias increases with the CV. If the model to be calibrated is a linear model with the exponent $k = 0$ or $k = 1$, $B(\hat{a}_k)$ decreases to 0, indicating that the sensitivity parameters of the linear terms in the model do not need to be adjusted because the variation of the scaling factor does not alter the calibration of the linear term parameters.

When the exponent is greater than 1, $B(\hat{a}_k)$ is always positive, as shown in Figure 2, implying that the calibrated sensitivity parameter of the corresponding non-linear term is always overestimated in the long run, and the calibrated parameter must be scaled down. For each fixed value of the CV with an exponent greater than 1, the extent of the embedded systematic bias increases with the exponent. If the exponent is smaller than 1 but greater than 0, the calibrated sensitivity parameters of the non-linear terms are expected to be underestimated in the long run because $B(\hat{a}_k)$ are evaluated to be negative.

3.3 Removal of systematic bias

In this subsection, a method for decreasing the systematic bias introduced by the derived global adjustment factor is proposed. The calibrated sensitivity parameter can be corrected by absorbing the variance of the scaling factor. To do so, the calibrated parameter must be divided by the global adjustment factor. The biased parameter, $\hat{a}_k$, associated with the k^{th} term with exponent k can be corrected as follows:

$$\bar{a}_k = \frac{\hat{a}_k}{\bar{F}_k} \tag{13}$$

where $\bar{a}_k$ is the globally corrected sensitivity parameter of the k^{th} term with exponent k.

4. Simulation

In this section, simulations are performed using sampled scaling factors from 100 lognormal distributions with different combinations of means and standard deviations, and hence different CVs, to demonstrate the correction power and efficiency of the derived global adjustment factor, $\bar{F}_k$. The association between the correction power and magnitudes of the mean and CV of the scaling factor is also investigated to illustrate the applicability of the global adjustment factor. Assuming that $a_0 = 3, a_1 = 0$, $a_2 = 1$, $n = 2$ and $m = 5$, the multivariate polynomial model chosen for the simulation is

$$y = 3 + X^2 = 3 + \left(\sum_{i=1}^{5} f_i x_i\right)^2 \tag{14}$$

where $X = \sum_{i=1}^{5} f_i x_i$.

4.1 Data generation

As the chosen $m = 5$, 5 sets of 10,000 observable independent variables, x_i, which serve as the observed data throughout the 100 simulations, are sampled from a negative exponential distribution with a mean of 0.2. In transport studies, many observable quantities are assumed to follow a negative exponential distribution. One example is the traffic flow of a particular fleet of vehicles, such as private cars and taxis. Due to the small volume of traffic flow of any particular type of vehicle at dawn, the distribution resembles a negative exponential. Thus, a negative exponential distribution is chosen to generate the data for the observable independent variable in this simulation. The value of $\frac{\overline{\sum_{i=1}^{m} x_i^2}}{\left(\sum_{i=1}^{m} x_i\right)^2}$ in this set of samples, x_i, is 0.333.

Because scaling factors are normally positive values, a lognormal distribution is chosen for the sampling. To formulate 100 simulations and examine the relationship between the correction power of the adjustment factor and the magnitude of the mean and CV of the scaling factor, scaling factors are sampled from 100 distributions with different combinations of means and standard deviations, such that both the mean and CV of the scaling factors range from 0.1 to 1.0 in steps of 0.1.

The corresponding 100 sets of dependent variables, y, serving as the observed dependent variables are then calculated based on the chosen polynomial using the assumed parameters, sampled scaling factors and observed independent variables.

4.2 Model calibration and parameter correction

Assuming that the means and variances are the only known information about the distributions of the scaling factors in these 100 simulations, then the observed independent variables can only be linearly projected by the corresponding mean of the scaling factors in each of the simulations. Regressions of the observed dependent variables on the linearly projected observed independent variables are performed to obtain 100 sets of calibrated parameters, $\hat{a}_0$ and $\hat{a}_2$.

Scaling factor variability is not expected to influence the calibration of parameter a_0, and a_2 is likely to be overestimated in the long run because the exponent of the non-linear term is greater than 1. The global adjustment factors, $\bar{F}_k$, for each simulation are evaluated from the exponent, the CV and $\frac{\overline{\sum_{i=1}^{m} x_i^2}}{\left(\sum_{i=1}^{m} x_i\right)^2}$ of the sampled independent variables, and applied to the calibrated $\hat{a}_2$ for correction.

4.3 Results

Because the mean value of $\hat{a}_0$ in the 100 simulations is 3.001(+0.1%), no adjustment is required, as expected. In contrast, the mean value of $\hat{a}_2$ is 1.127, which is 12.7% greater than the true value. The graph of the calibrated model shown in Figure 3 reveals a discrepancy between the true and calibrated models, and the distance between them increases with X. The mean value of the global adjustment factor, $\bar{F}_2$, is 1.128, which is greater than 1 as expected and suggests an overestimation of $\hat{a}_2$in the long run due to the loss of information about the scaling factor variance during calibration. This is consistent with the result obtained. The detected bias embedded in $\hat{a}_2$, $B(\hat{a}_2)$, is +12.8%. After adjustment, the mean value of the adjusted parameter $\bar{a}_2$ is 0.999, which is only 0.1% less than the true value. The graph of the corrected model illustrates that the derived global adjustment factor is capable of scaling down the overestimated calibrated model toward the true model. This demonstrates the significant correction power of the global adjustment factor.

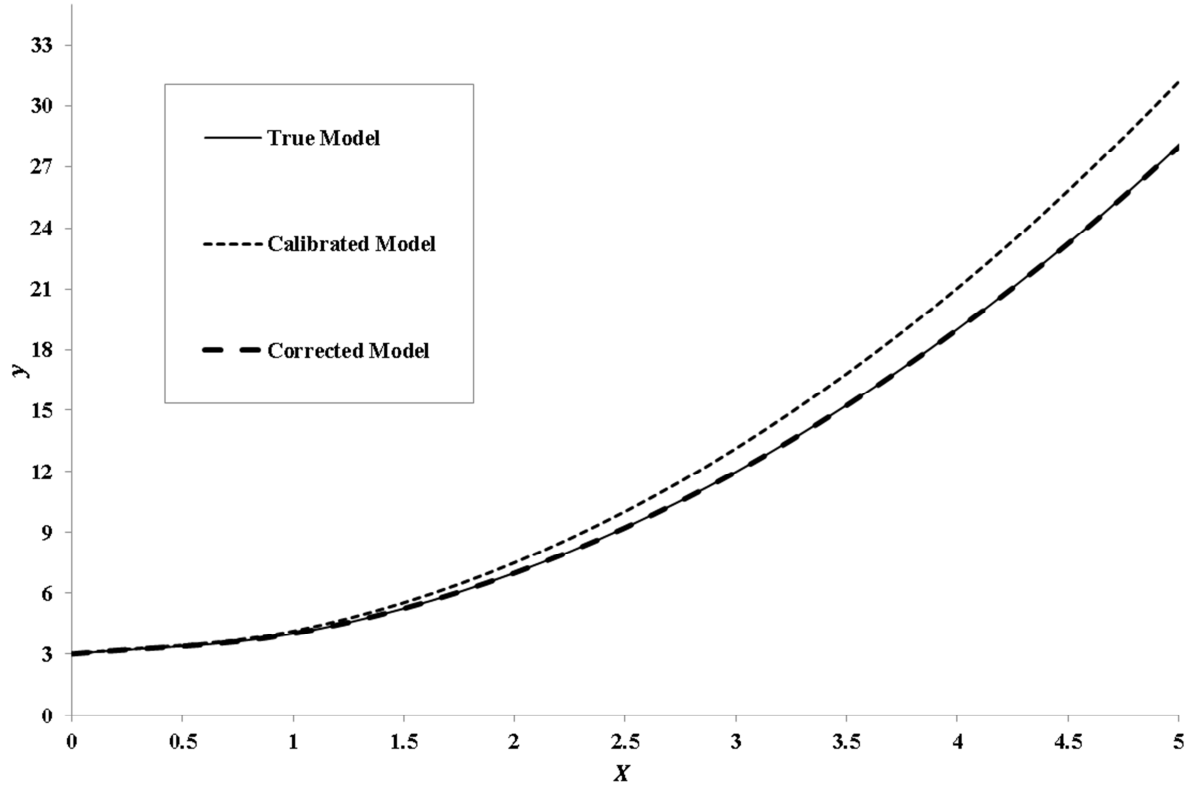

Figure 3. Demonstration of the correction power of the global adjustment factor.

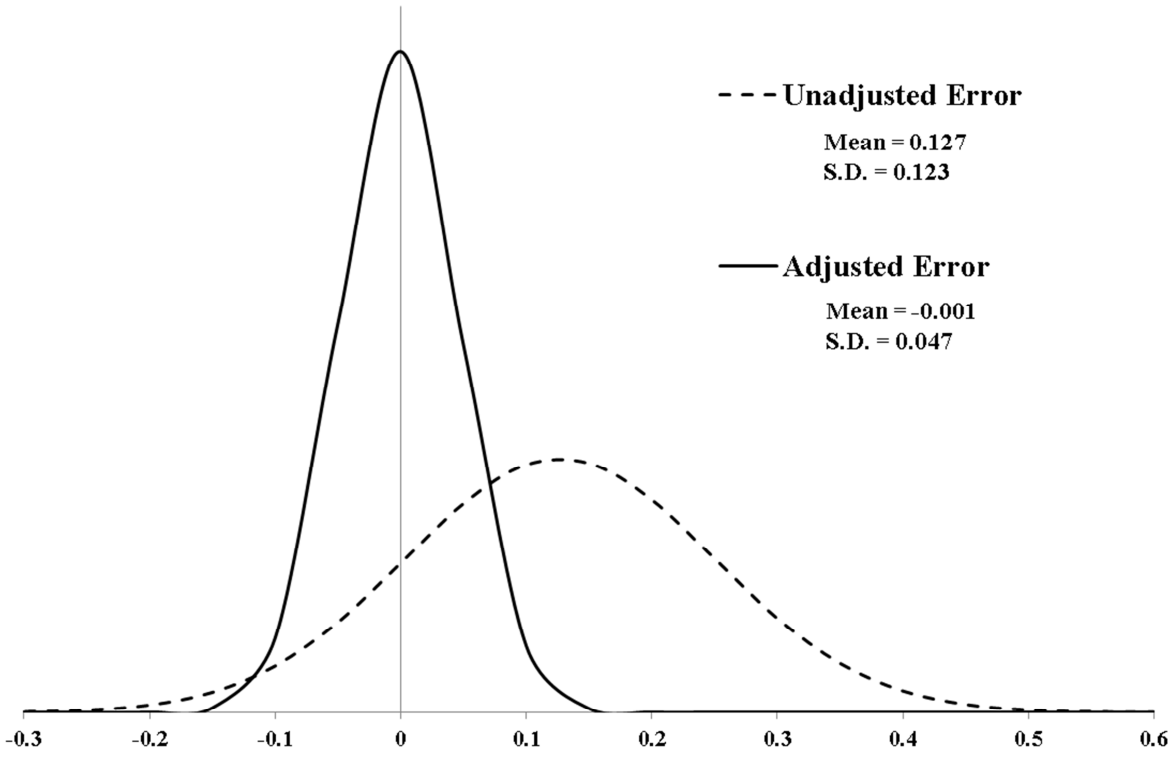

Figure 4. Distributions of the unadjusted and adjusted error in $\hat{a}_2$.

The mean and variance of the adjusted error can be used to assess the correction power of the global adjustment factor. The distribution of the error, centered at zero with a small standard deviation, represents the unbiased efficient estimation of the parameter in the long run because the central tendency has zero error and minimal dispersion. Figure 4 illustrates the correction power of the adjustment factor due to the shifting and narrowing of the distribution of the unadjusted errors. Before adjustment, the mean of the distribution of the unadjusted error in $\hat{a}_2$ deviates to the right of zero by 0.127. This provides an indication of the extent of the embedded systematic bias due to the linear data projection. Application of the adjustment factor shifts the distribution of the unadjusted error by 0.128, so that it deviates from zero by only -0.001. This shifting of the central tendency indicates the significant correction power of the adjustment factor. The narrowing of the distribution shows that the adjustment factor is also capable of confining the spread of the error. The correction decreases the standard deviation of the unadjusted error from 0.123 to 0.047. This reduction in the error spread demonstrates that the proposed global adjustment factor improves the efficiency of the parameter estimation in the long run.

4.4 Applicability of the global adjustment factor

The adjusted errors are regressed on the mean and CV of the scaling factor to investigate whether the applicability of the global adjustment factor is restricted to the magnitudes of each and to study the associations between them.

$$\varepsilon_{adj} = \beta_0 + \beta_1 \bar{f} + \beta_2 \frac{\sigma_f}{\bar{f}} \quad (15)$$

where ε_{adj} is the error of the calibrated parameter.

The calibrated parameters of the mean and CV of the scaling factor, $\hat{\beta}_1$ and $\hat{\beta}_2$, are -0.001 and 0.001, with p-values of 0.931 and 0.957, respectively. There is no evidence that the error is correlated with the mean and CV of the scaling factor. It is apparent that the applicability of the global adjustment factor is not likely to depend on these factors.

5. Case studies

To illustrate the application of the derived global adjustment factor for generalized multivariate polynomial model in Section 3, case studies using real-world data were conducted in relation to the model calibration of Macroscopic Bureau of Public Road (MBPR) function for six 1 km x 1 km regions in Tin Hau, Ma Tau Wai, Fortress Hill, Admiralty, Jordan and Kowloon Tong, Hong Kong. The MBPR function, which is in generalized multivariate polynomial form, is an essential input for the continuum modeling of urban cities (Ho and Wong, 2006, Ho et al., 2013, Wong, 1998, Yang and Wong, 2000, Yin et al., 2013).

5.1 Databases

These case studies modeled the travel time and total traffic flow relationship macroscopically using one-year travel time and traffic volume data associated with six selected 1x1 km regions obtained from the *Annual Traffic Census* (ATC) (Transport Department, 2010) and 480 GPS-equipped taxis.

The ATC provides detailed traffic data from over 1,500 stations covering 87% of trafficable roads in Hong Kong (Lam et al., 2003, Tong et al., 2003). The average annual daily traffic (AADT) across each of the stations where on-road fixed detectors are installed can be obtained from the ATC report.

The taxi GPS database stores detailed travel information about the 480 taxis over the course of 2010. Each of the 480 probe vehicles reported their real-time locations, expressed in terms of WGS84 (ITRF96 reference frame) in decimal degrees, and the dates, times, traveling directions and instantaneous speeds and occupancies to the traffic control center at a rate of twice per minute. Due to the full coverage of the taxi data over the entire road network, occupied taxi flow on any road segment can be easily counted. The travel time and taxi flow data were extracted from the 480 GPS-equipped taxis.

5.2 Necessity of linear data projection and adjustment of calibrated parameters

The MBPR function is a monotonically increasing non-linear model with a generalized multivariate polynomial form depicting the sensitivity of the travel time per unit distance to the increase in traffic volume associated with a defined studied region.

Hourly total traffic flow, which was calculated as the sum of hourly traffic flow entering the arbitrarily defined studied region through the roadways intercepting the defined boundaries, was the target independent variable. However, a non-negligible subset of urban network links was not outfitted with inductive loop detectors for direct and accurate measurements of traffic flow through these roadways, making the hourly total traffic flow entering the studied regions unobservable. As taxi GPS data were available with approximately full coverage, the occupied taxi flow on any street was readily attainable. A data scaling method such as linear data projection with a total-traffic-to-occupied-taxi ratio as the scaling factor could be leveraged to project the occupied taxi flows through the roadways diverting traffic into the studied regions as estimates of the hourly traffic flows.

Given the geographical proximity of the roadway within the network, the same set of vehicles should have more or less circulated within the network for each period. However, due to the heterogeneities of the road hierarchy and land use of different lots affecting the travel pattern, a homogenous traffic mix was not an entirely valid assumption. Thus, the traffic composition ratio across the network of the studied region could generally be assumed to follow a distribution subject to a certain level of spatial variability. With the availability of AADT data, ATC stations within the studied regions served as the sampling sites for the scaling factors. The distribution mean, which was the most likely traffic composition ratio, was estimated using the sampled scaling factors and used in the linear data projection.

As the scaling factors were subject to spatial variation and MBPR was a nonlinear function of the scaling factor, according to Proposition 1, systematic bias might have been introduced into the calibrated parameters due to the linear data projection. The variance of the scaling factor accounting for spatial variability was estimated using the sampled scaling factors. Furthermore, because the MBPR was a generalized multivariate polynomial, the derived global adjustment factor shown in Proposition 2 and proposed methodology were adopted to correct the systematically biased parameter and improve the reliability of the calibrated model.

5.3 Data extraction

The travel time and total traffic flow associated with the studied region were the essential ingredients in these case studies. Assuming that occupied taxis possess similar travel characteristics and behavior to those of other types of vehicles, only occupied taxis are considered in this paper. Figure 5 shows the normalized patterns of the hourly occupied taxi flows and hourly traffic flows at a few of the locations for which hourly counts were available within the studied regions.

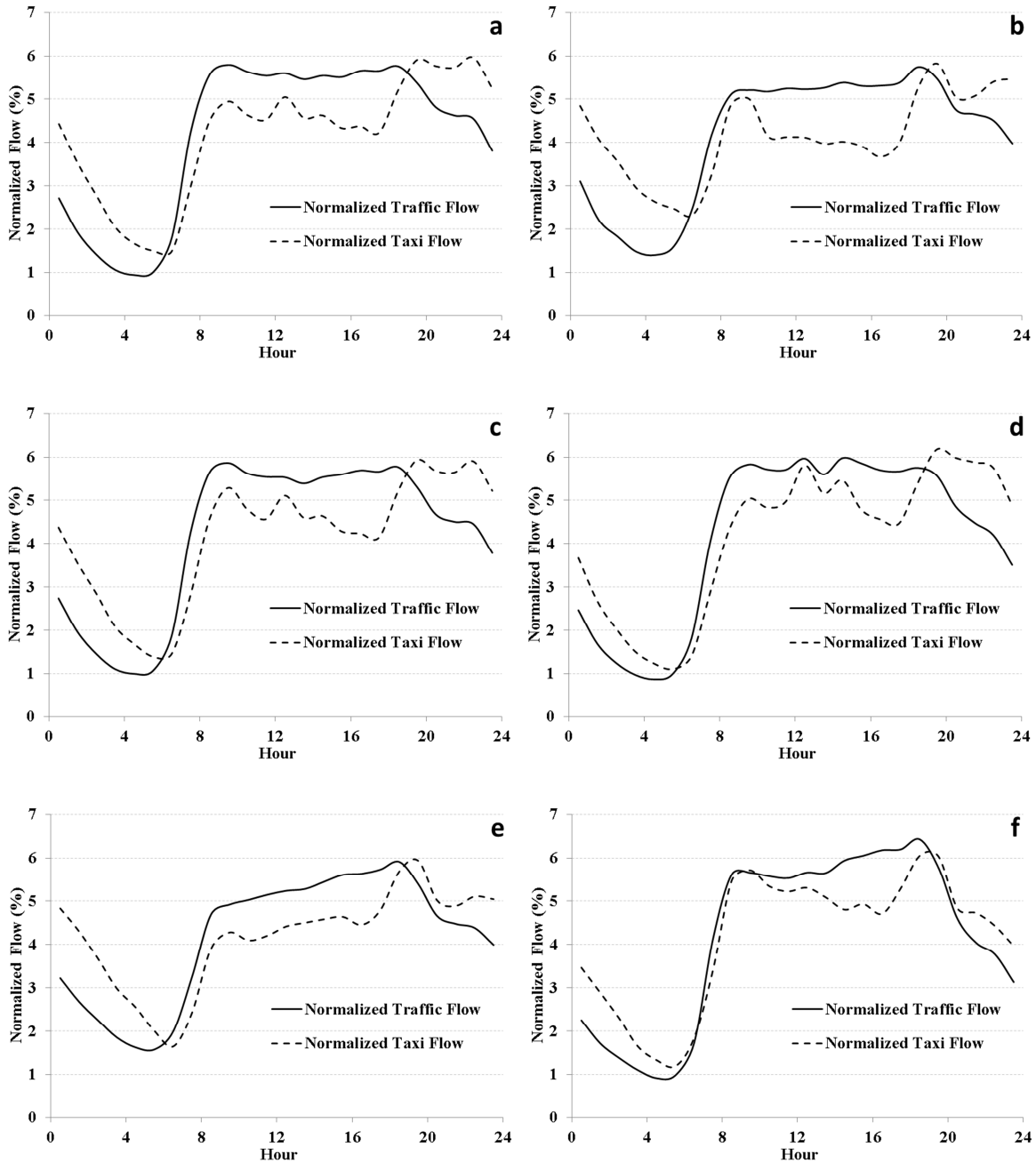

Figure 5. Normalized hourly occupied taxi flow and hourly total traffic flow patterns at several locations within the studied regions: (a) Tin Hau; (b) Ma Tau Wai; (c) Fortress Hill; (d) Admiralty; (e) Jordan; (f) Kowloon Tong.

Although both the normalized occupied taxi flows and normalized total traffic flows of the six studied regions varied throughout the day, their patterns remained remarkably similar, suggesting that the proposed assumption was reasonably valid. In other words, the hourly total traffic flows entering a studied region were inferred from the occupied taxi flows using the mean of the scaling factors.

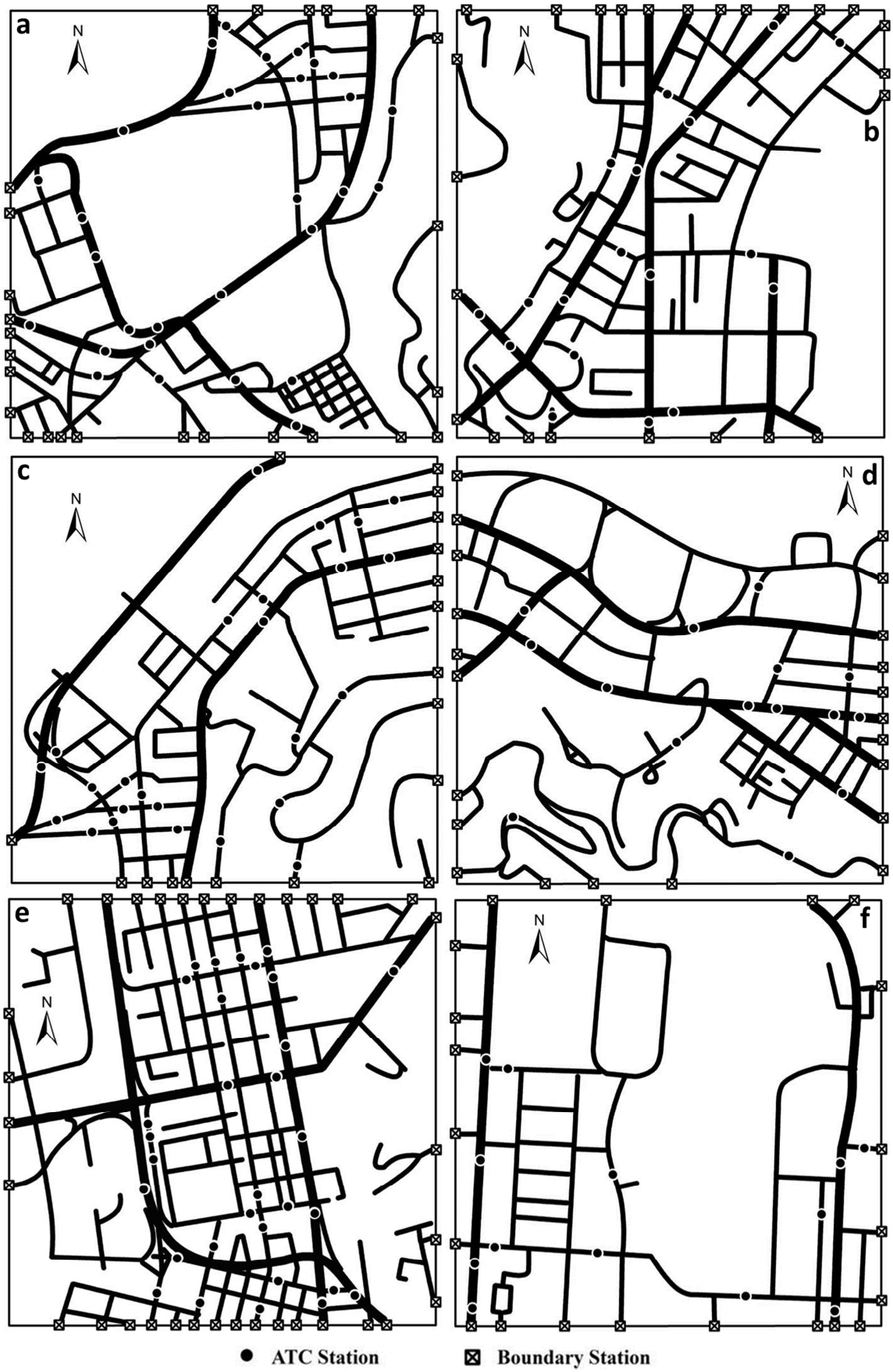

Figure 6. Schematic diagram showing the locations of ATC stations and boundary stations in the 1x1 km studied regions: (a) Tin Hau; (b) Ma Tau Wai; (c) Fortress Hill; (d) Admiralty; (e) Jordan; (f) Kowloon Tong.

Figure 6 is a schematic diagram showing the network skeletons of the six chosen studied regions, the locations of the ATC stations and the distributions of the boundary stations surrounding the 1x1 km boundaries. The ATC stations within each of the studied regions were the selected sampling sites for the scaling factors, which were the traffic composition ratios of total traffic to occupied taxi flows. The daily occupied taxi flows across each of the ATC stations were counted using the taxi GPS data. Dividing the AADT obtained from the ATC report by the average daily occupied taxi flow of each ATC station gave a sampled scaling factor. A mean $\bar{f}$ was an estimate of the average number of vehicles represented by each occupied taxi in that studied region. A standard deviation σ_f measured the spatial variation of the scaling factor across the network concerned. The means and standard deviations of the distributions of the scaling factors of the six studied regions were estimated from the sampled scaling factors, and are shown in the following table.

Table 1
Estimates of the mean and standard deviation of the scaling factor distributions for each studied region

Studied Region	Number of ATC stations	$\bar{f}$	σ_f
Tin Hau	31	193.4	74.0
Ma Tau Wai	18	193.9	81.9
Fortress Hill	25	232.2	93.8
Admiralty	15	211.2	71.4
Jordan	31	157.4	69.2
Kowloon Tong	13	193.4	45.4

Tin Hau, Ma Tau Wai and Kowloon Tong exhibited similar sampled scaling factor means. The traffic composition ratios were about 193, meaning that each occupied taxi roughly represented 193 vehicles in the three regions. Compared with other regions, occupied taxis comprised even smaller vehicular populations in Fortress Hill and Admiralty. Thus, each occupied taxi represented over 200 vehicles in these two regions. On the contrary, there were more occupied taxis in Jordan, with each occupied taxi representing about 157 vehicles. Because the scaling factors were sampled at different road segments and locations within the studied regions, the scaling factor standard deviation is a measure of the spatial variation of the traffic composition ratio that is the dispersion of the scaling factor across the network spatially. Among the six regions, the scaling factor standard deviation in Fortress Hill was the highest at 93.8, and the lowest was 45.4 in Kowloon Tong. Both the means and standard deviations of the scaling factors in each region could have resulted from factors such as the land use patterns, demography and household wealth of the corresponding regions.

For each of the studied regions, the hourly occupied taxi flows across the boundary stations were projected linearly using the corresponding scaling factor means as the estimates of the hourly traffic flows across these boundary stations. The total hourly traffic flow entering the studied region was the sum of these estimates.

Because the occupied taxis interacted with the surrounding vehicles as they travelled within the network of the studied regions, both types of vehicles should have travelled at similar speeds. Therefore, the speed of the occupied taxis was considered an unbiased estimate of the speed of the surrounding vehicles. The average instantaneous speed of all of the occupied taxis for each hour within the studied region was used as the estimate of the space-mean speed. The reciprocal of the space-mean speed was taken as the estimate of the average travel time per unit distance of movement, which served as the observed dependent variable. With both the collected travel time and linearly projected total hourly traffic flow taken into account, the MBPR function could be calibrated.

5.4 Model calibration and results

The MBPR function, which modeled the relationship between the travel cost and traffic flow associated with a region, was defined as follows:

$$T = T_f + T_f \alpha \left(\sum_{i=1}^{m} f_i v_i \right)^n \tag{16}$$

where T is the travel time; T_f is the free-flow travel time; m is the number of boundary stations; f_i is the scaling factor of boundary station i; v_i is the observed hourly taxi flow entering the studied region through a boundary station i, adjusted by the ratio between the two normalized patterns in Figure 5 during the observation hour; the sum of the products of f_i and v_i over m boundary stations, $\sum_{i=1}^{m} f_i v_i$, is the estimate of the total hourly traffic flow entering the studied region through these boundary stations; and α and n are the model parameters.

Several MBPR functions with different values of n, including 2, 3 and 4, were chosen as the candidate models. In general, the MBPR function in quadratic form was found to best fit the collected dataset. Thus, the MBPR function with n equal to 2 was chosen to study the effect of the spatial variability of the scaling factor on the extent of systematic bias embedded in the sensitivity parameter due to linear data projection. Models were calibrated on the observed travel time and linearly projected traffic flow. Table 2 shows the estimated parameters and R-squared values of each calibrated model in the studied regions.

The second column contains the calibrated free-flow travel time per unit distance, $\hat{T}_f$, of each studied region. Because free-flow travel time is the intercept of MBPR function, it is anticipated that no bias is embedded in it. The inverse of the calibrated free-flow travel time is an estimate of the free-flow travel speed, $\hat{v}_f$. The values in the third column are the calibrated sensitivity parameters of the nonlinear term in the MBPR function. They indicate that correction using the derived adjustment factor is necessary. As free-flow travel time is unbiased, the embedded systematic bias is absorbed by $\hat{\alpha}$. Both the free-flow travel time and congestion sensitivity parameter are deemed to be the aggregated result of the topological

features of the network. The last column, which contains the R-squared values, reveals the goodness-of-fit of the calibrated MBPR functions for the six selected studied regions.

Table 2
Model calibration results of the MBPR function in each of the studied regions

Studied Region	$\widehat{T}_f$(h/km)	$\widehat{v}_f$(km/h)	$\widehat{T}_f\widehat{\alpha}$(h^3/km/veh^2)	R^2
Tin Hau	0.0331	30.2	8.638E-11	0.483
Ma Tau Wai	0.0229	43.6	1.030E-10	0.715
Fortress Hill	0.0226	44.3	6.769E-11	0.600
Admiralty	0.0225	44.5	3.228E-11	0.690
Jordan	0.0356	28.1	1.724E-11	0.630
Kowloon Tong	0.0264	37.9	2.530E-10	0.729

Table 3
Adjustment of sensitivity parameter α

Studied Region	$\sigma_f/\bar{f}$	m	$\frac{\overline{\sum_{i=1}^{m} v_i^2}}{(\sum_{i=1}^{m} v_i)^2}$	$\bar{F}_2$	$B(\widehat{\alpha})$(%)	$\widehat{\alpha}$(h^2/veh^2)	$\bar{\alpha}$(h^2/veh^2)
Tin Hau	0.383	28	0.237	1.035	+3.5%	2.607E-09	2.519E-09
Ma Tau Wai	0.423	26	0.285	1.051	+5.1%	4.489E-09	4.271E-09
Fortress Hill	0.404	18	0.276	1.045	+4.5%	2.999E-09	2.870E-09
Admiralty	0.338	22	0.195	1.022	+2.2%	1.435E-09	1.404E-09
Jordan	0.440	35	0.116	1.022	+2.2%	4.837E-09	4.731E-09
Kowloon Tong	0.235	19	0.218	1.012	+1.2%	9.588E-09	9.474E-09

The CV of the scaling factor estimated by the sampled scaling factors of the ATC stations in each region and the mean of the sum of squared over squared of sum of the occupied taxi flow through each boundary station were required to evaluate the adjustment factor for the sensitivity parameter of the squared-term in the MBPR function. The CV of the scaling factor, defined as the ratio of the standard deviation to the mean, was a measure of the spatial variation of the scaling factor per unit vehicles, represented by each occupied taxi in this case. The mean of sum of squared over squared of sum of the occupied taxi flow was bounded by the upper bound of 1 and the lower bound evaluated from the reciprocal of the number of boundary stations, m. It was a measure of the average uniformity of the occupied taxi flow through the boundary stations over each hour throughout the year. It hit the lower bound when the occupied taxis entering the studied region were evenly distributed across the

boundary stations. It reached the upper bound if all of the occupied taxis entering the region were concentrated through one boundary station. Table 3 presents the adjustment factor, detected systematic bias and sensitivity parameter, α, before and after adjustment for each studied region.

The systematic biases detected in the calibrated sensitivity parameters of the squared terms of the six MBPR functions were positive, indicating that the calibrated sensitivity parameters were overestimated as expected. Due to the relatively high scaling factor variability (81.9) and the highest unevenly distributed occupied taxi flow through the boundary stations (0.285), the percentage of bias detected in the calibrated sensitivity parameter in the MBPR function of Ma Tau Wai was +5.1%, the highest of the studied regions. The high spatial variability of scaling factor and low uniformity of occupied taxi flow could have resulted from the heterogeneities of road hierarchy and land uses in different lots in the studied region. Although Fortress Hill showed the highest scaling factor variability (93.8), the highest scaling factor mean (232.2) resulted in a slightly lower CV (0.404) compared with that of Ma Tau Wai (0.423). Thus, the amount of bias embedded was +4.5%, lower than that of Ma Tau Wai. In addition to the lower spatial variability of the scaling factor (74.0), the more uniformly distributed occupied taxi flow through the boundary station (0.237) in Tin Hau led to the third highest embedment of systematic bias in the sensitivity parameter (+3.5%). The sensitivity parameters of the MBPR functions of Admiralty and Jordan shared the same extent of systematic bias (+2.2%). Although the CV of the scaling factor in Jordan (0.440) was greater than that in Admiralty (0.338), the substantial number of boundary stations in Jordan (35) significantly minimized the influence of the random effects of scaling factors and non-uniformity of the occupied taxi flows, and resulted in the lowest mean of sum of squared over squared of sum of the occupied taxi flow (0.116) out of all of the regions. This cancelled the effect of the high CV, and hence the embedded biases in the parameters of the models of Admiralty and Jordan were at the same level. Kowloon Tong is a residential area in Hong Kong with a relatively low density and a relatively homogenous road hierarchy and land use. This led to the lowest variability (45.4) and CV of scaling factor (0.235), and hence remarkably decreased the systematic bias embedded in the sensitivity parameter of the nonlinear term (+1.2%).

Given the evaluated adjustment factors, the systematic bias embedded in the calibrated sensitivity parameter was then corrected by the methodology proposed in Section 3.3. Figure 7 illustrates the scatter plots of travel time against projected vehicular flow, the calibrated MBPR functions and adjusted MBPR functions of the six 1x1 km studied regions in Hong Kong. The adjusted MBPR functions were the more reliable models describing the relationship between the travel time per unit distance and the traffic flow.

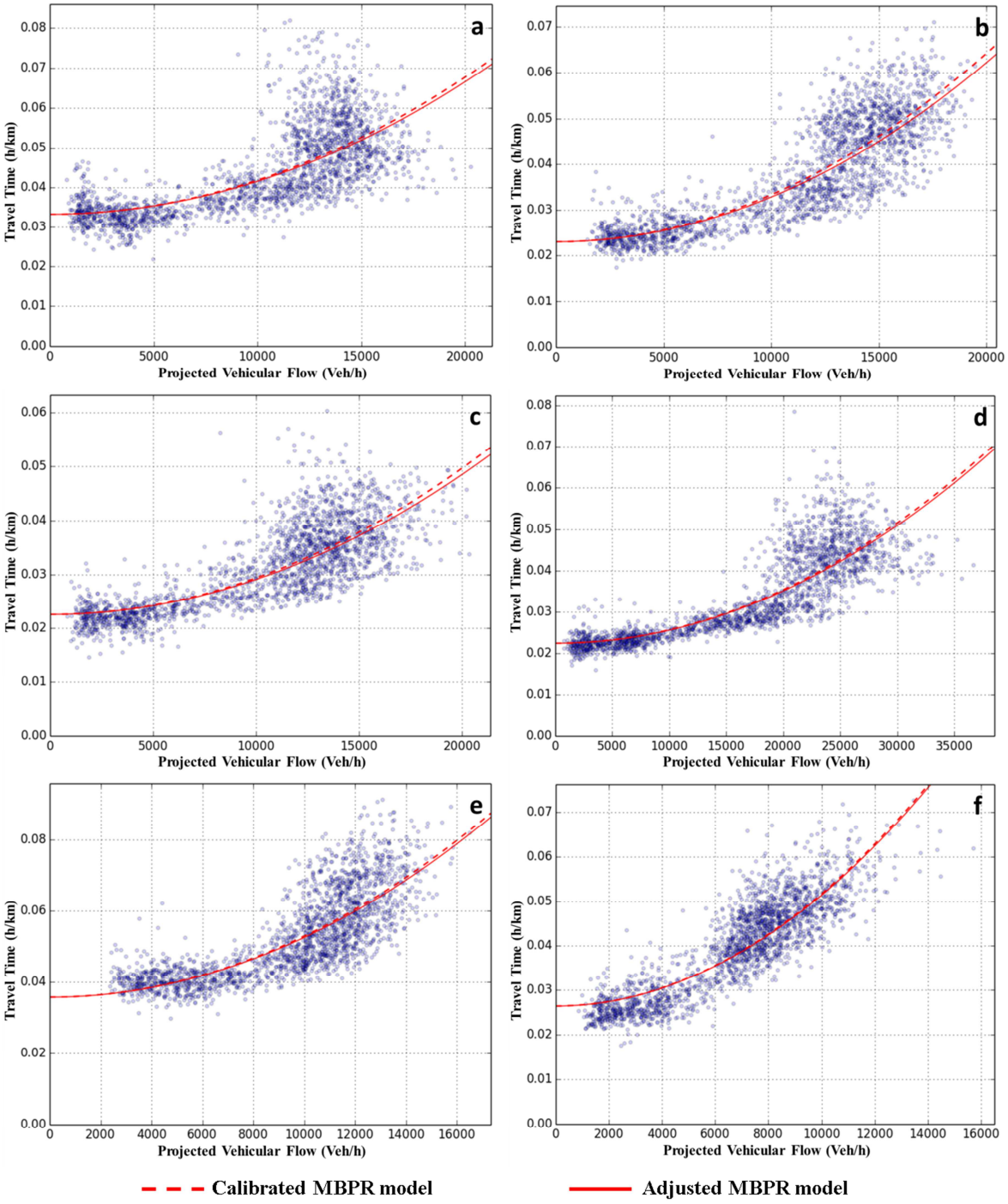

Figure 7. Scatter plots of travel time against vehicular flow, the calibrated MBPR functions and the adjusted MBPR functions for the 1x1 km studied regions: (a) Tin Hau; (b) Ma Tau Wai; (c) Fortress Hill; (d) Admiralty; (e) Jordan; (f) Kowloon Tong.

6 Conclusion

In the transportation field, using different instruments to acquire data representing population traffic characteristics through direct measurement may meet with various limitations and restrictions. Traffic data inferences are often made based on the data of a population subset. Linear data projection is a prevailing method adopted for data inference.

However, the possibility of a systematic bias being introduced into the parameters of models calibrated from linearly projected data has surprisingly gone unexplored. This paper unveils the necessary and sufficient condition of the introduction of systematic bias into calibrated parameters due to linear data projection. A systematic bias is introduced if the model to be calibrated is a nonlinear function of the scaling factor, and the scaling factor used is subject to variability. The embedded bias originates from systematically distorted data points caused by linear data projection in which the contribution of the scaling factor variance to the mean of the response variable undesirably remains in the error component.

In this paper, a generalized multivariate polynomial function model is applied to a discussion of the removal of systematic bias. Global adjustment factors are derived for each sensitivity parameter of the terms with different exponents. Both the metric detecting the extent of the embedment of systematic bias and method of its removal are subsequently proposed. The CV of the scaling factor, the exponent of the scaling factor and the mean of sum of squared over squared of sum of the observable independent variable are identified as the factors affecting the extent of the embedded systematic bias. If the exponent is greater than 1, then overestimation of the sensitivity parameter is anticipated. In contrast, underestimation is expected if the exponent is between 0 and 1. The amount of embedded systematic bias increases with both the CV of the scaling factor and the mean of sum of squared over squared of sum of the observable independent variable.

Comprehensive simulation has demonstrated the significant correction power and efficiency of the derived adjustment factor. There is no evidence to suggest that the applicability of the adjustment factor is restricted to the magnitudes of the mean and the CV of the scaling factor. This ensures and boosts confidence in the proposed approach to the removal of systematic bias. One of the major contributions of the proposed methodology is that the systematic bias introduced by data inference can be reduced via mathematic treatment without incurring additional equipment and installation costs at the data acquisition stage to ensure more accurate data.

Real-life traffic data from an on-road fixed detector and taxi GPS data were collected and integrated to illustrate how the derived adjustment factor can be applied in model calibrations and to calibrate MBPR functions in generalized multivariate polynomial form for six 1x1 km regions in Hong Kong. The extent of the embedded systematic bias was affected by the spatial variability of the scaling factor and the uniformity of the occupied taxi flow through the boundary stations. The derived adjustment factor successfully captured the information provided by the scaling factor variability undesirably remained in the error component. The adjusted MBPR functions were the more reliable models for depicting how the travel time per unit distance related to the traffic flow in these six regions.

The adjustment factor proposed in this paper is applicable for polynomial models. However, systematic bias is introduced into any model form as long as it is a nonlinear function of the scaling factor and when linear data projection is adopted. If the calibrated model is a factor of the second derivative of the model in relation to the scaling factor, then the biased parameter and effect of the scaling factor variability can be easily identified, as the

scaling factor variance can be grouped with the model parameter. The exponential function and some trigonometric functions such as the sine and cosine functions possess this property. Thus, the adjustment factors of the parameters of these models can be derived by following steps and procedures similar to those used to derive the adjustment factors for the sensitivity parameters of generalized multivariate polynomial models. For models that do not have this property, the effect of the ignorance of the scaling factor variability on parameters biased by linear data projection can be more complicated and ambiguous, as the variance of the scaling factor cannot be easily fused with the model parameters and it may be cumbersome to manipulate. Future studies could extend the techniques used to remove the systematic bias embedded in more complicated model forms as a result of linear data projection.

Acknowledgements

The work described in this paper was supported by a Research Postgraduate Studentship and grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU 17208614). We would like to express our sincere thanks to Concord Pacific Satellite Technologies Limited and Motion Power Media Limited for providing the taxi GPS data, and to the Transport Department of the HKSAR Government for providing the traffic flow data from the ATC.

References

Akcelik, R., 1978. A new look at Davidson's travel time function. *Traffic Engineering & Control*, 19 (10), 459-463.

Akcelik, R., 1980. Time-dependent expressions for delay, stop rate and queue length at traffic signals. Australian Road Research Board, Melbourne, Australia.

Ban, X. J., Li, Y., Skabardonis, A. & Margulici, J. D., 2010. Performance evaluation of travel-time estimation methods for real-time traffic applications. *Journal of Intelligent Transportation Systems,* 14 (2), 54-67.

Bertini, R. L. & Tantiyanugulchai, S., 2004. Transit buses as traffic probes: empirical evaluation using geo-location data. *Transportation Research Record: Journal of the Transportation Research Board*, 1870, 35-45.

Bhaskar, A. & Chung, E., 2013. Fundamental understanding on the use of Bluetooth scanner as a complementary transport data. *Transportation Research Part C: Emerging Technologies*, 37, 42-72.

Bolla, R. & Davoli, F., 2000. Road traffic estimation from location tracking data in the mobile cellular network. Proceedings of the Wireless Communications and Networking Conference, IEEE, Chicago, USA, 3, 1107-1112.

Caceres, N., Romero, L. M., Benitez, F. G. & Del Castillo, J. M., 2012. Traffic Flow Estimation Models Using Cellular Phone Data. *IEEE Transactions on Intelligent Transportation Systems,* 13 (3), 1430-1441.

Carlsson, R., Otto, A. & Hall, J. W., 2013. The role of infrastructure in macroeconomic growth theories. *Civil Engineering and Environmental Systems*, 30 (3-4), 263-273.

Chandra, S., Kumar, V. & Sikdar, P. K., 1995. Dynamic PCU and Estimation of Capacity of Urban Roads. *Indian Highways*, 23 (4), 17-28.

Davidson, K. B., 1966. A flow travel time relationship for use in transportation planning. Proceedings of the 3rd Australian Road Research Board (ARRB) Conference (Part 1), 183-194, Australian Road Research Board, Melbourne.

Del Castillo, J. M. & Benitez, F. G., 1995a. On the functional form of the speed-density relationship—I: general theory. *Transportation Research Part B: Methodological,* 29, 373-389.

Del Castillo, J. M. & Benitez, F. G., 1995b. On the functional form of the speed-density relationship—II: empirical investigation. *Transportation Research Part B: Methodological,* 29, 391-406.

Dowling, R. G., Singh, R. & Cheng, W. W.-K. 1998. Accuracy and performance of improved speed-flow curves. *Transportation Research Record: Journal of the Transportation Research Board,* 1646, 9-17.

Drake, J. S., Schofer, J. L. & May, A.D., 1967. A statistical analysis of speed density hypothesis. Highway Research Record 154, 53–87.

Drew, D. R., 1965. Deterministic aspects of freeway operations and control. Highway Research Record 99, 48–58.

García-ródenas, R., Verastegui-rayo, D., 2013. Adjustment of the link travel-time functions in traffic equilibrium assignment models. *Transportmetrica A: Transport Science,* 9 (9), 798-824.

Geroliminis, N. & Daganzo, C. F., 2008. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B Methodological,* 42 (9), 759-770.

Herrera, J. C. & Bayen, A. M., 2010. Incorporation of Lagrangian measurements in freeway traffic state estimation. *Transportation Research Part B: Methodological,* 44, 460-481.

Herrera, J. C., Work, D. B., Herring, R., Ban, X., Jacobson, Q. & Bayen, A. M., 2010. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. *Transportation Research Part C: Emerging Technologies,* 18, 568-583.

Herring, R., Hofleitner, A., Abbeel, P. & Bayen, A., 2010. Estimating arterial traffic conditions using sparse probe data. Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems, 929–936, Madeira Island, Portugal.

Ho, H. W. & Wong, S. C., 2006. Two-dimensional continuum modeling approach to transportation problems. *Journal of Transportation Systems Engineering and Information Technology,* 6 (6), 53-72.

Ho, H. W., Wong, S. C. & Sumalee, A., 2013. A congestion-pricing problem with a polycentric region and multi-class users: a continuum modelling approach. *Transportmetrica A: Transport Science,* 9 (6), 514-545.

Hofleitner, A., Herring, R. & Bayen, A., 2012. Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning. *Transportation Research Part B: Methodological,* 46 (9), 1097-1122.

Jayakrishnan, R., Tsai, W. K. & Chen, A., 1995. A dynamic traffic assignment model with traffic-flow relationships. *Transportation Research Part C: Emerging Technologies,* 3, 51-72.

Jenelius, E. & Koutsopoulos, H. N., 2013. Travel time estimation for urban road networks using low frequency probe vehicle data. *Transportation Research Part B: Methodological,* 53, 64-81.

Jie, L., van Zuylen, H., Chunhua, L. & Shoufeng, L., 2011. Monitoring travel times in an urban network using video, GPS and Bluetooth. *Procedia-Social and Behavioral Sciences,* 20, 630-637.

Kerner, B. S. & Konhäuser, P., 1994. Structure and parameters of clusters in traffic flow. *Physical Review E,* 50, 54-83.

Kwong, K., Kavaler, R., Rajagopal, R. & Varaiya, P., 2009. Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors. *Transportation Research Part C: Emerging Technologies,* 17 (6), 586-606.

Lam, W. H. K., Hung, W. T., Lo, H. K., Lo, H. P., Tong, C. O., Wong, S. C. & Yang, H., 2003. Advancement of the annual traffic census in Hong Kong. *Proceedings of the ICE-Transport,* 156, 103-115.

Lederman, R. & Wynter, L., 2011. Real-time traffic estimation using data expansion. *Transportation Research Part B: Methodological,* 45 (7), 1062-1079.

Liang, Y., Reyes, M. L. & Lee, J. D., 2007. Real-time detection of driver cognitive distraction using support vector machines. *IEEE Transactions on Intelligent Transportation Systems,* 8 (2), 340-350.

Lum, K. M., Fan, H. S. L., Lam, S. H. & Olszewski, P., 1998. Speed-flow modeling of arterial roads in Singapore. *Journal of transportation engineering,* 124 (3), 213-222.

Macnicholas, M. J., 2008. A simple and pragmatic representation of traffic flow. Symposium on The Fundamental Diagram: 75 years, Transportation Research Board, Woods Hole, MA.

Miwa, T., Ishiguro, Y., Yamamoto, T. & Morikawa, T., 2013. Allocation planning for probe taxi devices based on information reliability. Transportation Research Part C: Emerging Technologies, 34, 55-69.

Moore, J. E., Cho, S., Basu, A. & Mezger, D. B., 2001. Use of Los Angeles Freeway Service Patrol Vehicles as Probe Vehicles. Technical report, Berkeley, CA.

Munjal, P. K. & Pipes, L. A., 1971. Propagation of on-ramp density perturbations on unidirectional two-and three-lane freeways. *Transportation Research,* 5 (4), 241-255.

Nanthawichit, C., Nakatsuji, T. & Suzuki, H., 2003. Application of probe-vehicle data for real-time traffic-state estimation and short-term travel-time prediction on a freeway. *Transportation Research Record: Journal of the Transportation Research Board,* 1855, 49- 59.

Peer, S., Knockaert, J., Koster, P., Tseng, Y.-Y. & Verhoef, E. T., 2013. Door-to-door travel times in RP departure time choice models: An approximation method using GPS data. *Transportation Research Part B: Methodological,* 58, 134-150.

Pipes, L. A., 1967. Car following models and the fundamental diagram of road traffic. *Transportation Research,* 1 (1), 21-29.

Schwarzenegger, A., Bonner, D. E., Iwasaki, R. H. & Copp, R., 2008. State Highway Congestion Monitoring Program (HICOMP), Annual Data Compilation. Technical report, Caltrans, Sacramento, CA.

Spiess, H., 1990. Conical volume-delay functions. *Transportation Science,* 24 (2), 153-158.

Tisato, P., 1991. Suggestions for an improved Davidson travel time function, *Australian Road Research,* 21 (2), 85-100.

Tong, C. O., Hung, W. T., Lam, W. H. K., Lo, H. K., Lo, H. P., Wong, S. C. & Yang, H., 2003. A new survey methodology for the annual traffic census in Hong Kong. *Traffic engineering & control,* 44 (6), 214-218.

Transport Department 2010. The annual traffic census 2010. Traffic and Transport Survey Division, Transport Department, Government of the Hong Kong SAR.

Transportation Research Board 2000. Highway capacity manual. National Research Council, Washington, DC.

Van Aerde, M. Single regime speed-flow-density relationship for congested and uncongested highways. Paper presented at the 74th Annual Meeting of the Transportation Research Board, Washington, DC, 1995.

Wong, S. C., 1998. Multi-commodity traffic assignment by continuum approximation of network flow with variable demand. *Transportation Research Part B: Methodological,* 32 (8), 567-581.

Wong, R.C.P., Szeto, W.Y., Wong, S.C. & Yang H., 2014. Modelling multi-period customer-searching behaviour of taxi drivers. *Transportmetrica B,* 2 (1), 40-59.

Wright, J. & Dahlgren, J., 2001. *Using vehicles equipped with toll tags as probes for providing travel times*, California PATH Program, Institute of Transportation Studies, University of California, and Berkeley, CA.

Yang, H. & Wong, S.C., 2000. A continuous equilibrium model for estimating market areas of competitive facilities with elastic demand and market externality. *Transportation Science,* 34 (2), 216-227.

Ygnace, J. L. & Drane, C., 2001, Cellular telecommunication and transportation convergence: a case study of a research conducted in California and in France on cellular positioning techniques and transportation issues. Proceedings of Intelligent Transportation Systems Conference, Oakland, CA, 16-22.

Yin, J., Wong, S. C., Sze, N. N. & Ho, H. W., 2013. A continuum model for housing allocation and transportation emission problems in a polycentric city. *International Journal of Sustainable Transportation,* 7 (4), 275-298.

Zhan, X., Hasan, S., Ukkusuri, S. V. & Kamga, C., 2013. Urban link travel time estimation using large-scale taxi data with partial information. *Transportation Research Part C: Emerging Technologies*, 33, 37-49.

Zhao, Y., 2000. Mobile phone location determination and its impact on intelligent transportation systems. *Intelligent Transportation Systems, IEEE Transactions on*, 1 (1), 55-64.

Zheng, F. & Van Zuylen, H., 2013. Urban link travel time estimation based on sparse probe vehicle data. *Transportation Research Part C: Emerging Technologies*, 31, 145-157.